A systematic evaluation and benchmarking of text summarization methods for biomedical literature: From word-frequency methods to language models

Fabio Baumgärtel [1,*], Enrico Bono [1,2,*], Lucas Fillinger [1], Louiza Galou [1,2], Kinga Keska-Izworska [1], Samuel Walter [1], Peter Andorfer [1], Klaus Kratochwill [1,2], Paul Perco [1,3], and Matthias Ley [1,2,✉]

[1]Delta4 GmbH, Vienna, Austria; [2]Division of Pediatric Nephrology and Gastroenterology, Department of Pediatrics and Adolescent Medicine, Comprehensive Center for Pediatrics, Medical University Vienna, Vienna, Austria; [3]Department of Internal Medicine IV, Medical University Innsbruck, Innsbruck, Austria; *These authors contributed equally to this work.

Abstract

The rapid expansion of biomedical literature demands automated summarization tools that can reliably condense research articles into concise, accurate overviews. We benchmarked 62 text summarization methods – ranging from frequency-based and TextRank extractors to modern encoder-decoder models (EDMs) and large language models (LLMs) – on a set of 1,000 biomedical abstracts for which author-generated highlights sections were available as reference summaries. Models were evaluated using a composite suite of metrics covering lexical overlap (ROUGE-1/2/L, BLEU, METEOR), embedding-based semantic similarity (RoBERTa, DeBERTa, all-mpnet-base-v2), and factual consistency (AlignScore). Our results indicate that general-purpose language models (LMs) achieve the highest overall scores across both lexical and semantic metrics, outperforming both reasoning-oriented and domain-specific models. Within the general-purpose group, medium-sized models, typically runnable on a single node, often outperform frontier-scale counterparts, suggesting an optimal balance between model capacity and computational efficiency. Statistical extractive methods lag behind all neural approaches. These findings provide a systematic reference for selecting summarization tools in biomedical research and highlight that broad pretraining remains more effective than narrow domain adaptation for generating high-quality scientific summaries.

biomedical text summarization | natural language processing | large language models benchmarking

Correspondence: ***matthias.ley@delta4.ai***

1 Introduction

The exponential growth of scientific literature has created a demand for text summarization methods to support scientists in efficiently prioritizing papers, extracting relevant information, and interpreting research findings. Automatic text summarization (ATS) methods have evolved from statistical approaches to deep learning-based models thus becoming increasingly sophisticated and reliable at capturing essential parts from complex research articles. Although ATS methods have been previously evaluated and described [1, 2], only a few have focused on scientific literature [3, 4] with no systematic benchmarking of ATS methods for biomedical literature being available.

1.1 Statistical and Encoder-Decoder Approaches

The pre-neural era of text summarization was mainly characterized by extractive approaches, where in an unsupervised way, summaries were generated by using word or concept frequencies to identify relevant sentences. The first word-frequency based approaches were discussed by Luhn [5], who presented a method based on the assumption that recurrent words in a text are likely more important. Later, Edmundson [6] introduced concepts such as cue words, title words, and sentence position to further enhance the automatic summarization process. The concept of term frequency-inverse document frequency (TF-IDF) was only adopted later [7] and applied to text summarization by representing sentences as term-weight vectors that down-weight frequently occurring terms with low context specificity in an entire corpus of documents, while promoting rarer terms that at the same time are more context-specific. Word-frequency based approaches have been extensively used in scientific text summarization and form the basis for a number of newer more sophisticated methods [8]. Finally, graph-based statistical methods, in which sentences are represented as nodes and their relationships as edges weighted by similarity measures (e.g. cosine similarity of TF-IDF vectors), allow for the identification of the most central information based on the documents global structure rather than local word counts. TextRank is one widely used representative in the biomedical domain, in which the PageRank algorithm is applied on a constructed sentence graph to compute sentence importance scores for generating text summaries based on top-ranked sentences [9].

With the advent of sequence-to-sequence (Seq2seq) frameworks, text summarization shifted toward neural approaches that paraphrase and condense text using encoder-decoder architectures, originally implemented

with recurrent neural networkss (RNNs), long short-term memory (LSTM) networks, and gated recurrent units (GRUs) [10, 11]. The introduction of self-attention mechanisms replaced Seq2seq frameworks by processing sequences in parallel rather than sequentially, enabling the capture of complex linguistic patterns and long-range contextual relationships [12]. This innovation laid the foundation for transformer architectures that quickly gained popularity in performing a wide range of natural language processing (NLP) tasks, including text summarization. One of the earliest and most influential transformer-based models, bidirectional encoder representations from transformers (BERT) [13], was widely adopted in domain specific tasks as it could be fine-tuned by adding a task-specific output layer. Inspired by BERT's architecture, several abstractive summarization models emerged, including bidirectional and auto-regressive transformer (BART) - a denoising autoencoder for pretraining Seq2seq models [14] that can be trained or fine-tuned on scientific literature [15, 16]. The text-to-text transfer transformer (T5) model was introduced as a unified text-to-text framework for a broad spectrum of NLP tasks due to its high flexibility with no need for architectural changes [17]. Pre-training with extracted gap-sentences for abstractive summarization sequence-to-sequence (PEGASUS) was specifically proposed for abstractive summarization [18] and has been adapted for scientific text with domain-specific variants including "google/pegasus-pubmed" and "google/bigbird-pegasus-large-pubmed". Robustly optimized BERT approach (RoBERTa) is an optimized version of BERT trained on a bigger corpus of text, which led to the creation of longformer [19], a transformer-based architecture that can handle longer texts for text-to-text generation, with "allenai/led-base-16384" and "led-large-16384-arxiv" as notable examples [20].

1.2 Language Models

Despite these advances, the field of ATS quickly moved toward decoder-only architectures which are at the basis of LLMs, able to capture semantic relations with higher flexibility and specificity. LLMs can be classified as (i) general-purpose models, which leverage their broad domain knowledge across diverse NLP tasks, (ii) reasoning-oriented models, characterized by logical text understanding through iterative chain-of-thought processing and instruction tuning [21], and (iii) domain-specific models, tailored for specialized tasks or specific scientific domains. Several families of LLMs have been developed, including the generative pre-trained transformer (GPT) series developed by OpenAI (GPT-1 [22] through GPT-5) and open-source variants like GPT:OSS, all pre-trained on large-scale text corpora through self-supervised learning. Similarly, Anthropic's Claude models are built on transformer architecture and trained through a constitutional AI approach [23]. This family also includes a series of reasoning models such as Sonnet-4 and Opus-4. Meta's large language model Meta AI (Llama) family [24], with LLaMA 3.1 as the most capable open-source model available to date, includes domain-specific adaptations such as OpenBioLLM-LLaMA-3 [25], a biomedical variant trained on a large corpus of high-quality biomedical data, and MedLLaMA-2 [26], a medical LM based on LLaMA 2 architecture. Google developed a series of lightweight models including the Gemma series [27], with Gemma3 as its latest and most powerful reasoning model. Microsoft introduced the Phi series [28, 29], which comprises Phi-4-reasoning and Phi-4-mini-reasoning, alongside BioGPT [30], a domain-specific model built on the GPT architecture and fine-tuned for biomedical applications. IBM released the Granite series [31], with Granite 4.0 as its reasoning-capable variant. Mistral AI developed the Mistral family [32], including Magistral as its first reasoning model [33], and Biomistral [34], an open-source variant pre-trained on PubMed Central data for biomedical text processing. Alibaba Cloud introduced the Qwen 3 series [35] as an open-source LLM family, which inspired SciLitLLM [36], a specialized model for scientific literature understanding based on Qwen2.5 and trained through continual pre-training (CPT) and supervised fine-tuning (SFT) on scientific literature [36]. DeepSeek has developed reinforcement learning (RL)-driven reasoning models that achieve performance comparable to state-of-the-art closed-source models while requiring only a fraction of their training costs [37]. Lastly, Apertus [38] represents Switzerland's first large-scale open, multilingual LM with a fully documented and openly accessible development process.

To the best of our knowledge, no comprehensive benchmarking of text summarization models on biomedical literature has been performed so far. This study addresses this gap by systematically evaluating 62 text summarization models, ranging from word-frequency methods to state-of-the-art LLMs, using a dataset of 1,000 biomedical abstracts and corresponding highlights sections as reference summaries for benchmarking. We provide actionable insights for selecting appropriate summarization tools to accelerate knowledge discovery in biomedical sciences and discuss strengths and limitations of the different approaches.

2 Materials and Methods

Figure 1 provides an overview of the complete benchmarking workflow of this study from dataset construction and model inference to metric computation, score aggregation, and statistical analysis.

2.1 Gold-Standard Dataset

We generated a gold-standard benchmarking dataset comprising 1,000 biomedical peer-reviewed articles from *ScienceDirect* and *Cell Press* - as these publishers provide standardized highlights sections for publications

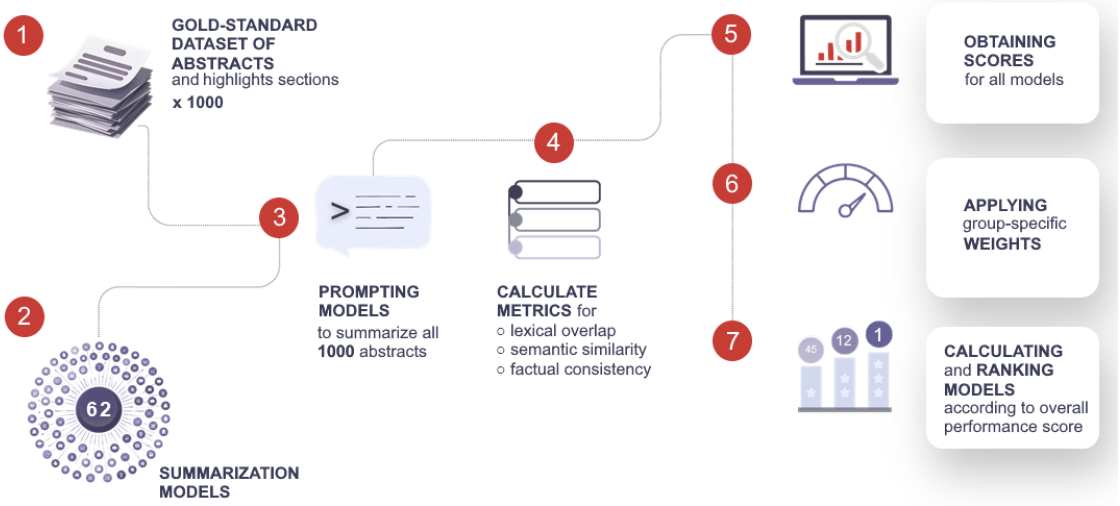

Figure 1. Overview of the benchmarking workflow: (1) Construction of a gold-standard dataset comprising 1,000 biomedical abstracts paired with author-provided highlights sections. (2) Suite of 62 diverse summarization models. (3) Uniform prompting of all models to generate summaries for each abstract. (4–5) Computation of summary quality metrics capturing lexical overlap, semantic similarity, and factual consistency. (6) Aggregation of metrics using group-specific weights to obtain overall performance scores. (7) Ranking of models based on overall performance scores.

[39, 40]. These highlights sections provide concise bullet points capturing the main findings of scientific articles. The concatenated highlights served as reference summaries in our evaluation, while the corresponding abstracts were used as input texts for the summarization task. While summarization is inherently subjective, author-generated highlights represent the most credible and standardized source, ensuring the captured content reflects the study's intended key messages.

Articles were collected systematically across a variety of journals from the two publishers to ensure coverage of different fields within molecular sciences, including among others drug discovery, genomics, proteomics, biotechnology, and biochemistry. We selected 50 articles from each of 20 different journals across the two publishers, resulting in 1,000 papers as shown in Table 1.

This setup provided standardized pairs of abstracts and reference summaries subsequently used in our systematic evaluation and benchmarking of ATS methods.

2.2 Summarization Methods

We in total evaluated 62 summarization models, ranging from simple frequency-based algorithms to small language models (SLMs) & LLMs. These models were classified into five categories, as listed in Table 2. We obtained the pre-trained EDMs through the Hugging Face library, selecting architectures widely used for abstractive summarization tasks to represent well established neural approaches within our benchmark. Additionally, we evaluated a range of popular SLMs and LLMs, defining SLMs as models with fewer than 10 billion parameters [41]. Both SLMs and LLMs with advanced reasoning capabilites were categorized as "reasoning-oriented" to investigate how multi-step problem solving affects summarization performance. Similarly, models fine-tuned on scientific/biomedical data or specifically tailored for text summarization were classified as "domain-specific" to assess the impact of domain adaptation. Overall, this selection covers various model sizes and release dates, ensuring a representative mix of both, widely adopted and recent, architectures.

Exceptionally large models, such as LLaMA 3.1 405B, were excluded as their computational requirements exceed those of practical summarization pipelines. Each of the 62 models was tasked with generating summaries for the 1,000 abstracts in the dataset, resulting in a total of 62,000 generated summaries for evaluation.

2.3 Prompt Design

To ensure comparability across models, we prompted all summarization systems with an identical task description. The prompt instructed the models to generate concise summaries focused on the main findings of each publication while excluding unnecessary background or methodological details. Each model received the publication title and abstract as an input and was asked to produce an output of 15–100 words. If the abstract did not contain any substantive results or conclusions, the model was instructed to return the predefined token INSUFFICIENT_FINDINGS.

The exact prompt used for all models was as follows:

```
Summarize the provided publication (title
and abstract) in 15-100 words.
Key requirements:
- Identify main findings, results, or
contributions
- Preserve essential context and nuance
- Exclude background, methods unless crucial
to conclusions
- Write concisely and objectively
- Avoid repetition and unnecessary
qualifiers
```

Table 1. Overview of journals included in the gold-standard dataset. Each journal contributed 50 articles, resulting in 500 articles from *ScienceDirect* and 500 articles from *Cell Press*.

Publisher	Journals
ScienceDirect	Drug Discovery Today; Journal of Molecular Biology; FEBS Letters; Journal of Biotechnology; Gene; Genomics; Journal of Proteomics; The International Journal of Biochemistry & Cell Biology; Cytokine; Developmental Cell
Cell Press	Cell; Cancer Cell; Cell Chemical Biology; Cell Genomics; Cell Host & Microbe; Cell Metabolism; Cell Reports; Cell Reports Medicine; Cell Stem Cell; Cell Systems

Table 2. Overview of summarization methods/models evaluated in this study, organized by category.

Category	Methods/Models
Traditional models	textrank; frequency
General-purpose EDMs	facebook/bart-base; google-t5/t5-base; google-t5/t5-large; google/pegasus-large
Domain-specific EDMs	facebook/bart-large-cnn; google/pegasusxsum; google/pegasus-cnn_dailymail; google/pegasus-pubmed; google/bigbird-pegasuslarge-pubmed; csebuetnlp/mT5_-multilingual_XLSum; led_large_16384_arxiv_summarization
General-purpose SLMs	gemma3:270M; gemma3:1b; gemma3:4b; PetrosStav/gemma3-tools:4b; granite3.3:2b; granite3.3:8b; granite4:tiny-h; granite4:small-h; granite4:micro; granite4:micro-h; llama3.1:8b; llama3.2:1b; llama3.2:3b; mistral:7b; phi3:3.8b; gpt-4o-mini; gpt-4.1-mini; chat_swiss-ai/Apertus-8B-Instruct-2509
General-purpose LLMs	gemma3:12b; mistral-nemo:12b; mistral-small3.2:24b; mistral-small-2506; mistral-medium-2505; mistral-large-2411; phi4:14b; gpt-3.5-turbo; gpt-4o; gpt-4.1; claude-3-5-haiku-20241022
Reasoning-oriented SLMs	deepseek-r1:1.5b; deepseek-r1:7b; deepseek-r1:8b; qwen3:4b; qwen3:8b;
Reasoning-oriented LLMs	deepseek-r1:14b; gpt-oss:20b; gpt-5-nano-2025-08-07; gpt-5-mini-2025-08-07; gpt-5-2025-08-07; claude-sonnet-4-20250514; claude-opus-4-20250514; claude-opus-4-1-20250805; magistral-medium-2509
Domain-specific SLMs	completion_microsoft/biogpt; medllama2:7b; chat_aaditya/OpenBioLLM-Llama3-8B; conversational_BioMistral/BioMistral-7B; chat_Uni-SMART/SciLitLLM1.5-7B
Domain-specific LLMs	chat_Uni-SMART/SciLitLLM1.5-14B

```
If no substantial findings exist, respond:
'INSUFFICIENT_FINDINGS'
```

2.4 Evaluation Metrics

As there is no single metric that can fully reflect summary quality, especially in the biomedical field where both coverage of key information and factual correctness are critical, we employed both surface-level metrics based on lexical overlap, referred to as lexical-based metrics, and embedding-level metrics. The latter includes metrics based on semantic similarity, denoted as semantic-based metrics, as well as one metric that evaluates factual consistency.

2.4.1 Surface-level Metrics. Surface-level metrics compare the generated summaries with the reference summaries mainly at the word or phrase level. While they do not capture meaning beyond surface overlap, they remain common metrics in summarization research and provide a straightforward foundation for evaluation. We used three recall-oriented understudy for gisting evaluation (ROUGE) variants (ROUGE-1, ROUGE-2, ROUGE-L) [42], bilingual evaluation understudy (BLEU) [43], and metric for evaluation of translation with explicit ordering (METEOR) [44]. ROUGE-1 and ROUGE-2 measure how many unigrams (single words) or bigrams (word pairs) from the reference appear in the generated output, while ROUGE-L identifies the longest sequence of words shared between the two. BLEU calculates how many n-grams in the output also occur in the reference, emphasizing precision over recall and applying a brevity penalty to counteract the tendency toward overly short summaries. METEOR extends n-gram matching by considering word stems and synonyms, making it more robust to wording variations. Together, these metrics offer a simple but transparent point of reference.

2.4.2 Embedding-based Metrics. To capture similarity beyond surface-level word overlap, we included a set of embedding-based metrics built on pre-trained transformer models. These methods generate vector representations of text, allowing them to capture semantic similarity rather than just word overlap. We employed RoBERTa [45] and decoding-enhanced BERT with dis-

entangled attention (DeBERTa) [46], two transformer-based models with strong performance across NLP tasks. In summarization evaluation, they can assess whether two summaries capture the same content even if phrased differently.

We further included all-mpnet-base-v2 [47], a transformer model fine-tuned for sentence similarity. Unlike RoBERTa and DeBERTa, which are general-purpose encoders, Masked and Permuted Pre-training (MPNet) was trained with a focus on alignment at the sentence-level. This characteristic makes it a useful complement to the other metrics, as it is particularly sensitive to whether the overall meaning of a reference summary is preserved in the system output.

Finally, to evaluate factual consistency, we applied AlignScore [48], a metric designed to assess whether the statements in a generated summary are supported by the source text. In contrast to the other metrics, AlignScore compares the output to the input text itself (i.e. the publication abstract) rather than the reference summary (i.e. the highlights section), as factual accuracy can only be assessed relative to the original input text. This addition ensures that our evaluation captures errors and hallucinations that might otherwise be overlooked.

2.4.3 Overall Performance Metric. To comprehensively assess the performance of each model on the summarization task, we employed a multi-metric framework covering three dimensions: lexical (n=5), semantic (n=3), and factual (n=1). To prevent the impact of dimension imbalance, we first computed an average score for each category as follows:

$$\text{avg}_{\text{lex}} = \frac{\text{ROUGE-1} + \text{ROUGE-2} + \text{ROUGE-L} + \text{METEOR} + \text{BLEU}}{5} \quad (1)$$

$$\text{avg}_{\text{sem}} = \frac{\text{RoBERTa} + \text{DeBERTa} + \text{all-mpnet-base-v2}}{3} \quad (2)$$

$$\text{avg}_{\text{fact}} = \text{AlignScore} \quad (3)$$

To examine how the different evaluation metrics relate to each other, we computed pairwise Spearman correlation coefficients across all models (Figure 2). Averaging metrics within each dimension ensures that highly correlated measures such as ROUGE-1, ROUGE-2, and ROUGE-L do not dominate the overall evaluation. Averaged scores were used to construct the overall performance score, weighting each dimension according to literature-based evidences. Semantic metrics for example correlate more strongly with human judgment than lexical ones and thus were given a slightly higher overall weight [49]. Furthermore, we accounted for the fact that AlignScore, due to its different point of reference, tends to favor extractive approaches [50], while showing only a moderate correlation with both lexical and semantic metrics (Figure 2). Based on these considerations, we defined an Overall Performance Score (OPS) as a weighted combination of the three dimension-wise average scores:

$$\text{OPS} = 0.35 \times \text{avg}_{\text{lex}} + 0.40 \times \text{avg}_{\text{sem}} + 0.25 \times \text{avg}_{\text{fact}} \quad (4)$$

We finally ranked models based on their overall performance scores, constructing the performance rank, with lower ranks indicating better performance. To examine which model performed best for each dimension (lexical, semantic, factual), we also ranked models based on average scores for each dimension.

2.5 Statistical Analysis

To assess differences in overall performance between model categories and families, we first applied Welch's Analysis of Variance (ANOVA) and conducted pairwise post-hoc comparisons using the Games-Howell test. This procedure is well suited for groups with unequal sample sizes and heterogeneous variances, conditions that apply to our benchmark due to the heterogeneity of models within each category and family. Adjusted p-values were used to determine the statistical significance of between-group differences.

2.6 Benchmarking Framework

The benchmark was conducted using Python 3.12. Gold standard data were retrieved from open-access articles published by *ScienceDirect* and *Cell Press* through manual extraction of titles, abstracts, and highlights sections, along with metadata including publication URLs, identifiers, section types, and article types where available. All data were stored in machine-readable JSON format.

The framework was implemented using the Python standard library supplemented by several specialized packages: pandas [51] for data import and export, scikit-learn [52] for computing cosine similarities of embeddings and TF-IDF vectors, networkx [53] for graph construction and PageRank algorithm [54]. Additional evaluation metrics were computed using NLTK [55] for METEOR and BLEU scores, ROUGE-score, BERT-score [56], AlignScore, and sentence-transformers [57] with the all-mpnet-base-v2 model.

Communication with proprietary closed-source LLMs was facilitated through the official Python APIs provided by Anthropic, Mistral AI, and OpenAI. Local LLM execution was performed on a workstation equipped with a NVIDIA RTX A4000 GPU (16GB VRAM) running Ollama as a backend service, accessed through its Python API along with the transformers library [58].

All LLMs were configured with a temperature parameter of 0.2 to optimize reproducibility while avoiding completely deterministic outputs. For the latest generation of OpenAI models featuring adaptive reasoning capabilities, the configuration was set to `text.verbosity = low` and `reasoning.effort = minimal`. The full set of parameters and prompts are documented in the `config.py` file in the GitHub repository.

2.7 Data Availability

The complete source code, documentation, gold standard dataset, and processed results are available at: https://www.github.com/Delta4AI/LLMTextSummarizationBenchmark.

3 Results

Our benchmark results offer a comparative view of summarization performance across all evaluated models on biomedical abstracts. We first examine the correlations among the chosen evaluation metrics and, based on the findings, identify the best-performing model across lexical, semantic, and factual dimensions. Next, we compare model performance across categories and families to identify significant differences. Finally, we present a case study illustrating how concept coverage varies between models.

3.1 Metric Correlations

Pairwise Spearman correlations were computed to analyze the relationships between the different evaluation metrics (Figure 2). Strong positive correlations were observed among most lexical-based metrics (ROUGE-1/2/L, METEOR, and BLEU), with the correlation between METEOR and BLEU marking an exception ($\rho = 0.23$). ROUGE variants showed almost identical behavior ($\rho > 0.89$), while BLEU and METEOR demonstrated slightly weaker but still substantial alignment with ROUGE measures ($\rho = 0.53 - 0.79$).

Most semantic-based metrics (RoBERTa, DeBERTa, and all-mpnet-base-v2) showed high internal consistency ($\rho > 0.5$), reflecting their shared focus on semantic similarity. When compared with lexical-based metrics, correlations were moderate to strong in most cases, indicating that both categories capture related but not identical dimensions of summary quality.

In addition, AlignScore, a factual consistency metric, correlated moderately with the semantic-based metrics even though they both are embedding-based ($\rho = 0.35 - 0.5$), as well as with the lexical-based ones ($\rho = 0.2 - 0.41$), which can be attributed to its different point of reference.

Overall, these relationships demonstrate that the various metrics are broadly consistent while providing complementary perspectives. This supports the use of an aggregated “Overall Performance Score” as a balanced indicator of overall summarization performance.

3.2 Overall Model Performance

Based on overall performance ranks, referred to as "Performance Rank" and derived from our multi-metric evaluation framework (2.4.3 Overall Performance Metric), models from the Mistral family were top-ranked with high performance scores across the three evaluated metric dimensions, as depicted in Figure 3.

The mistral:7b model ranked first, followed by mistral-small3.2:24b and mistral-small-2506. The lowest-ranked models included bigbird-pegasus-large-pubmed and pegasus-pubmed from the PEGASUS family, and mT5_multilingual_XLSum from the T5 family. Overall, domain-specific SLMs and EDM models showed poor performance across all the different dimensions of metrics.

Among the 10 top-ranked models, six were general-purpose LLMs and four were general-purpose SLMs. A similar trend was observed in the top half of the ranking (positions 1 to 31) with the presence of some reasoning-oriented LLMs. In contrast, in the lower half of the ranking, where models start to perform poorly across most metric dimensions, the majority were reasoning-oriented SLMs, general-purpose EDMs, domain-specific EDMs, and traditional models.

Considering each metric dimension separately, some Mistral models ranked high in the lexical dimension with GPT-5-nano, a reasoning-oriented LLM from the GPT family, ranked third. In the semantic dimension, Mistral models achieved high ranks, but the highest positions were occupied by the Phi4:14b general-purpose LLM and some models from the Granite family, such as granite4:tiny-h and granite3.3:2b. Finally, the highest ranks based on the factual dimension were covered by traditional and encoder-decoder models. Detailed performance indications for each model across all individual ranked metrics are provided in supplementary Figure S1.

3.3 Category Comparisons

To compare model performance across categories, we displayed the distribution of overall performance scores within each category, along with the best- and worst-performing individual models (Figure 4a). Results of the Games-Howell post-hoc comparisons between individual groups are given in Figure 4b.

All general-purpose LLMs achieved high overall scores thus forming the top category of models with general-purpose SLMs following second however showing a slightly wider spread in performance scores. Reasoning-oriented LLMs and domain-specific LLMs followed in third and fourth place based on average overall performance.

Domain-specific EDMs, domain-specific SLMs, traditional models, and general-purpose EDMs performed significantly worse as compared to general-purpose LMs. Domain-specific EDMs and domain-specific SLMs displayed the widest spread of performance as can be seen in Figure 4a.

3.4 Family Comparisons

To complement the category-level findings, we next examined performance across model families, since models within the same family often share architectural features or training strategies that may cause consistent performance trends. Figure 5a shows the distribution of overall performance scores across all model families. Similar to the category-level analysis, there

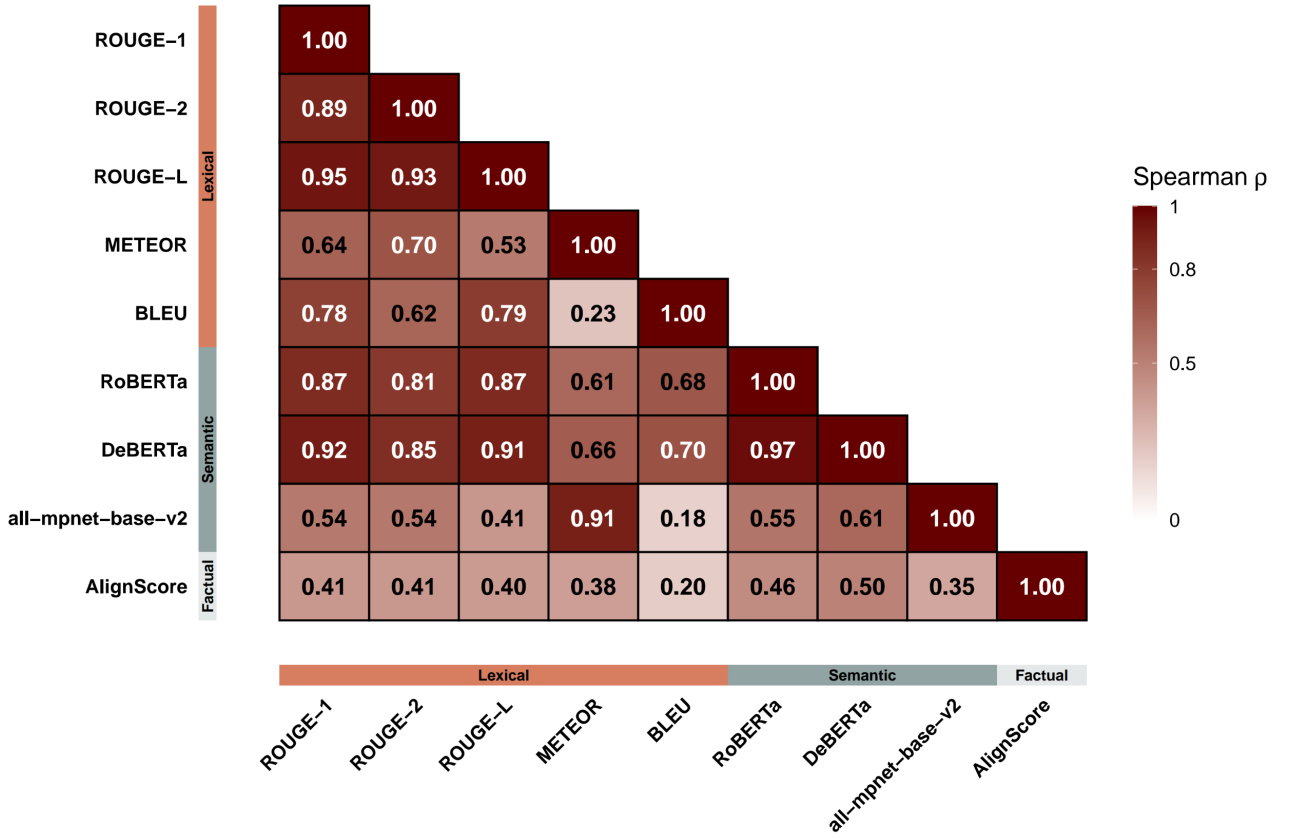

Figure 2. Spearman Correlation between Metrics: Correlation matrix of evaluation metrics. Spearman correlation coefficients (ρ) between two metrics based on their mean scores across all models are given. Metric categories (lexical, semantic, factual) are indicated on the x and y axes. For visualization purposes, cells with values higher than 0.8 are shown with white text.

were clear performance differences between architectural lineages. The Mistral family achieved the strongest overall results, with several models ranking among the highest-scoring models in the entire benchmark. Families dominated by modern LLM or SLM architectures, such as Granite, Claude, and Gemma, consistently outperformed more traditional extractive approaches and families built on encoder-decoder architectures such as longformer encoder-decoder (LED), BART, T5 and PEGASUS.

In particular, the GPT and Llama families showed lower performance scores than expected given the competitive performance of their top models. This was primarily due to the inclusion of domain-specific small models, such as BioGPT and OpenBioLLM-Llama3-8B, which performed substantially worse than their general-purpose counterparts and therefore pulled down the family-level averages.

Regarding the Games-Howell post-hoc matrix in the family comparison (Figure 5b), a combination of significant and non-significant differences between families was observed, reflecting the greater heterogeneity within some lineages. While many high-performing families differed significantly from traditional and encoder-decoder dominated families, several comparisons among modern SLMs and LLMs dominated families did not reach statistical significance.

3.5 Qualitative Analysis: Case Study

To illustrate how concept coverage differs between models beyond aggregate metrics, we examine two summaries of a biomedical research article [59] against publisher-provided highlights. The source article, titled *"A percolation phase transition controls complement protein coating of surfaces"*, includes four key highlights: (H1) The complement protein network has a switch-like response when attacking surfaces; (H2) Complement "decides" to coat surfaces if surface protein spacing is below a threshold; (H3) Complement's threshold decision-making arises from a percolation phase transition; (H4) Complexity science shows how complement makes discrete decisions attacking surfaces.

We selected mistral-small3.2:24b as a high-performing model (ranked 2nd overall) and medllama2:7b as a substantially lower-ranked model (ranked 54th out of 62) to contrast differences in concept coverage. The generated summaries are as follows:

- **mistral-small3.2:24b:** "This study reveals that the complement system's activation on materials exhibits a sharp threshold behavior, controlled by the surface density of complement attachment points. This percolation-like phase transition occurs across scales, from nanoparticles to macroscale materials, and is driven by a minimal subnetwork of complement proteins. The findings

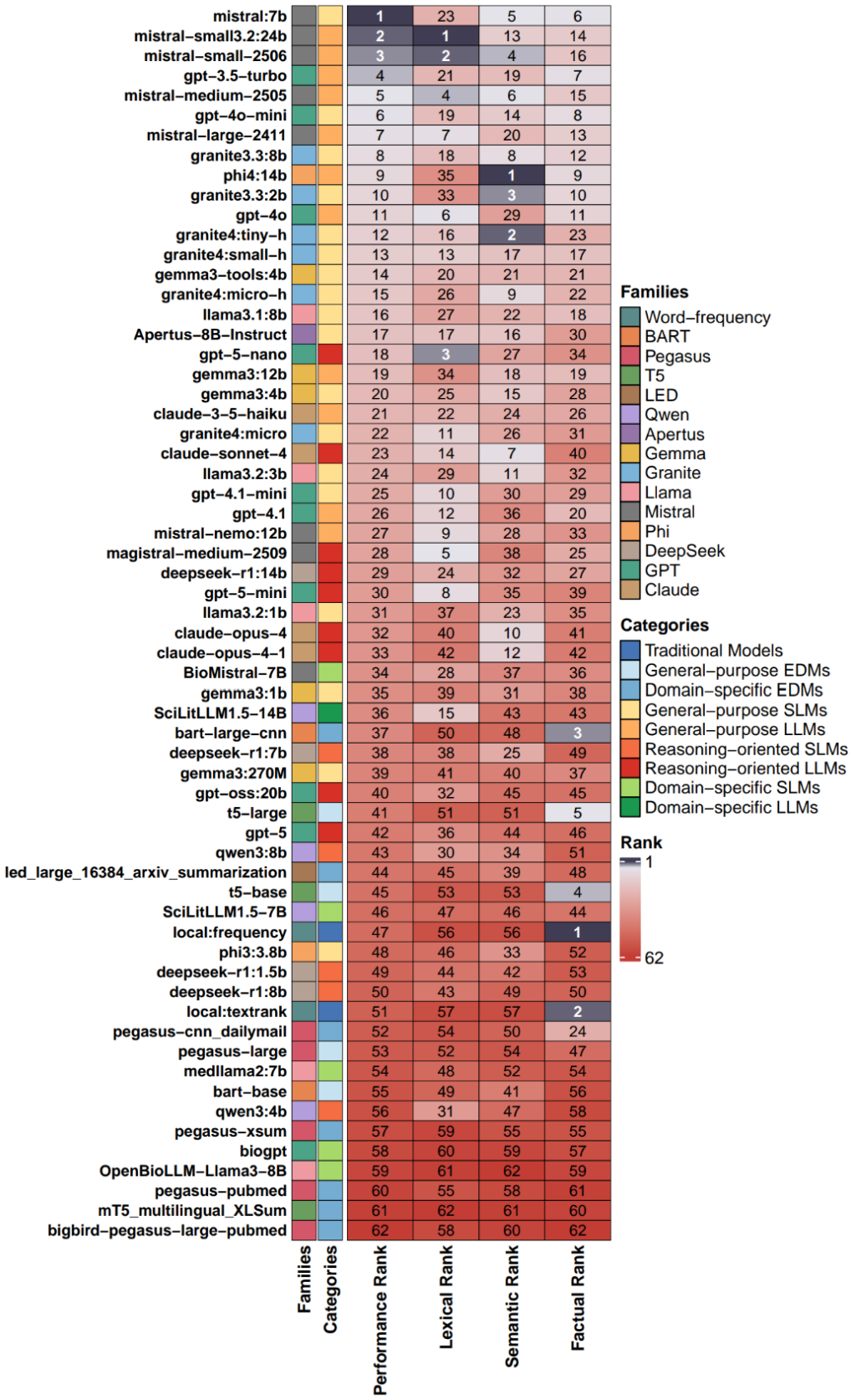

Figure 3. Ranking of Summarization Models: Overview of the performance of all evaluated models across lexical, semantic and factual metric dimensions. Each row corresponds to one model. Columns show the overall Performance Rank and independent ranks for each metric dimension, with lower ranks indicating better performance. The figure displays each model's family and category. Models are sorted by the overall Performance Rank. Ranks are displayed using a continuous diverging color gradient, ranging from navy-blue tones for top-performing models through light neutral shades for mid-ranked models to dark red tones for low-performing models. Text color is switched to white for the highest-ranked models (ranks ≤ 3) to improve readability.

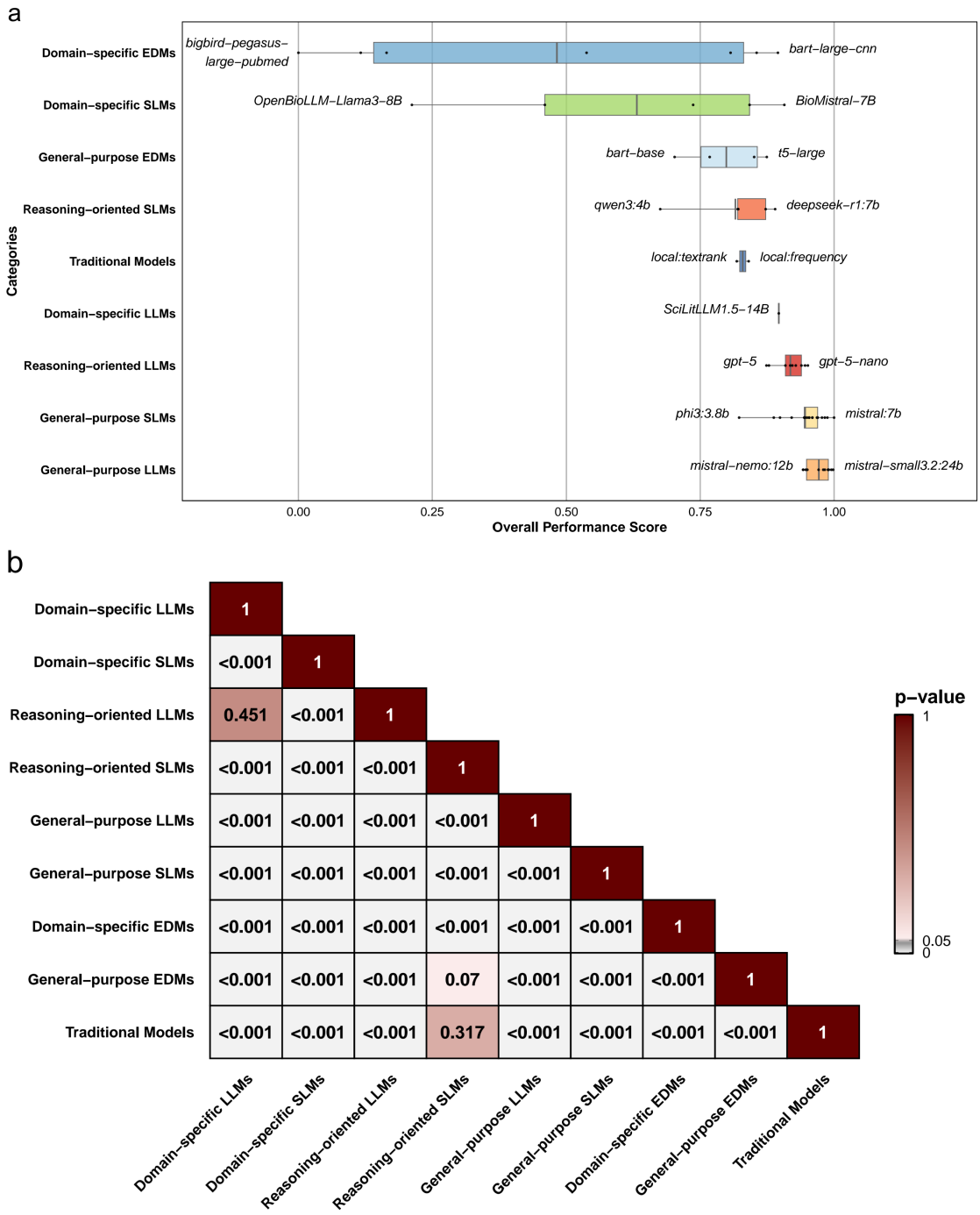

Figure 4. (a) Overall performance of Model Categories: Distributions of overall performance scores across the nine model categories, with each model represented by a black dot. Categories are ordered from lowest to highest-performing according to the median performance score. The best- and worst-performing models of each category are highlighted. **(b) Games-Howell Post-hoc Test across Model Categories:** Games–Howell pairwise adjusted p-value matrix comparing all category pairs. Each cell shows the adjusted p-value for the difference in metric mean scores between two categories, with white-to-gray shading indicating smaller p-values.

explain how complex protein networks 'decide' to interact with and attack materials, initiating inflammation."

- **medllama2:7b:** "The main findings are that there is a sharp threshold for complement activation upon varying a fundamental material parameter and this threshold manifests at scales spanning single nanoparticles to macroscale pathologies. The study shows how a minimal subnetwork of complement proteins can give rise to percolation-type critical transitions in the complement response, which explains why some materials are more likely to trigger an immune response than others."

Table 3 summarizes the concept-level coverage and overall performance for both models. While both mod-

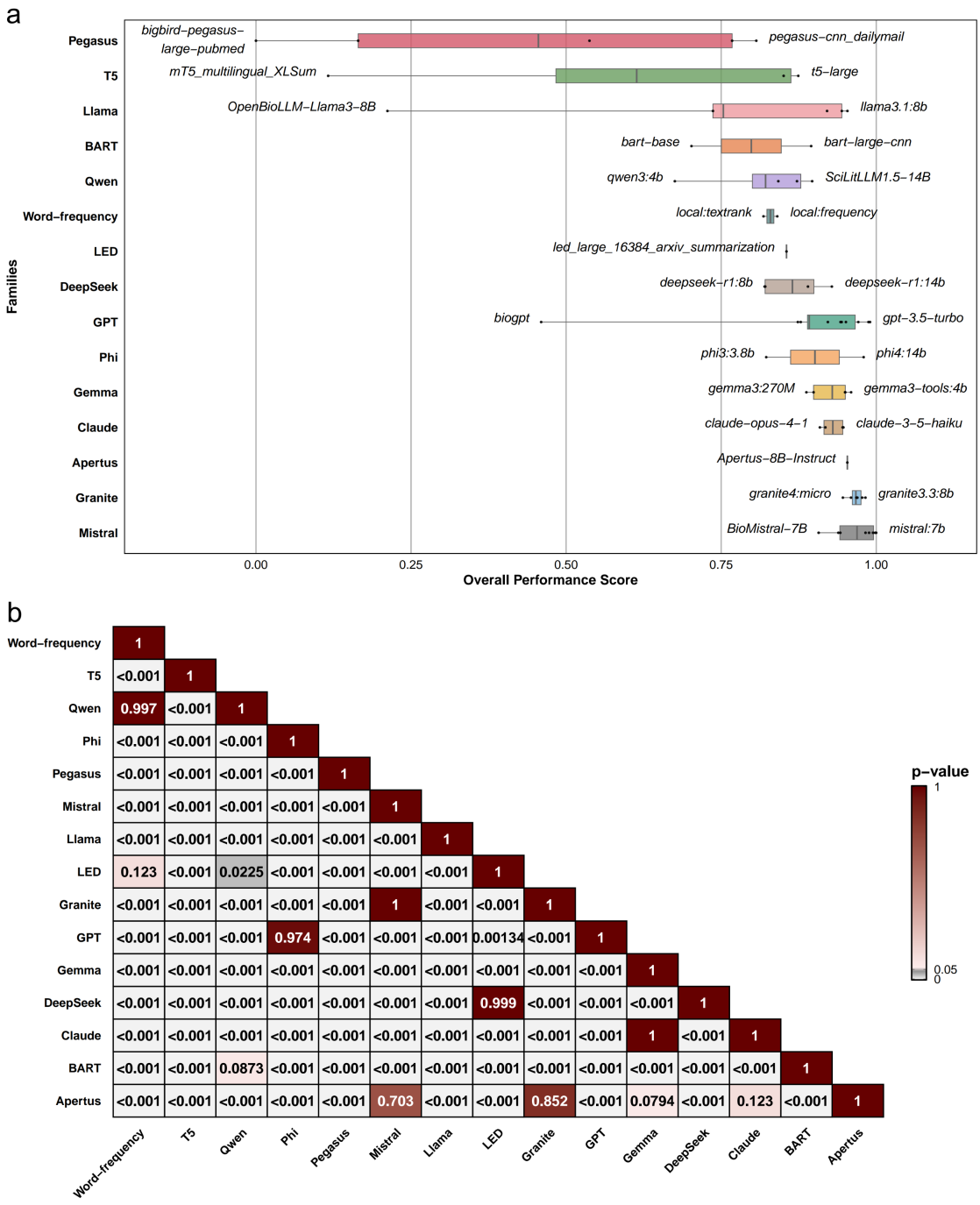

Figure 5. a) Overall performance of Model Families: Distributions of overall performance scores across all model families, with each model represented by a black dot. Families are ordered from lowest to highest-performing according to the performance score. The best- and worst-performing models of each family are highlighted. **b) Games-Howell Post-hoc Test across Model Families:** Games–Howell pairwise adjusted p-value matrix comparing all family pairs. Each cell shows the adjusted p-value for the difference in metric mean scores between two families, with white-to-gray shading indicating smaller p-values.

els capture the switch-like behavior (H1) and the percolation transition (H3), the lower-performing model fails to explicitly link the threshold to surface density/spacing (H2) and omits the complexity-based decision-making framework (H4). In contrast, mistral-small3.2:24b achieves full coverage by explicitly connecting the physical density to the "decisions" made by the complex protein network. The overall performance scores, derived using the weighted metric aggregation introduced in Section 2.4.3, further reflect this difference, with mistral-small3.2:24b achieving a higher score compared to medllama2:7b.

4 Discussion

Our benchmarking analysis of 62 different text summarization models on a dataset of 1,000 abstracts revealed clear performance differences. General-purpose LLMs achieved the highest summarization quality across all metric dimensions closely followed by general-

Table 3. Concept coverage analysis of model-generated summaries. Symbols: ✓ = fully covered; ∼ = partially covered; × = not covered. Overall performance scores are derived using the weighted metric aggregation described in Section 2.4.3.

Reference Concept	mistral-small3.2:24b	medllama2:7b
H1: Switch-like response	✓	✓
H2: Surface density threshold	✓	∼
H3: Percolation phase transition	✓	✓
H4: Discrete decision-making	✓	×
Coverage Score	4.0 / 4.0	2.5 / 4.0
Overall Performance Score	0.659	0.512

purpose SLMs and reasoning-oriented LLMs. In contrast, domain-specific models, encoder–decoder architectures, and traditional extractive methods performed significantly regarding overall performance. These results highlight the clear progression from extractive and encoder–decoder approaches toward transformer-based models, while also showing that domain-specific fine-tuning alone does not necessarily lead to improved summarization quality.

Comparing models by architecture, size, and domain focus shows that, overall, LLMs perform best, likely due to their high number of parameters which enable them to better understand the complex context typical for biomedical literature. While very small models like Gemma3:270M can indeed lack the capacity to handle this complexity, SLMs remain competitive, with some models, such as mistral:7b and gpt-4o-mini, even outperforming certain LLMs across the three metric dimensions. This may be attributed to the fact that smaller datasets are often more curated and of higher quality compared to the large amount of data required to train a large model [60].

Interestingly, medium-sized models (e.g many models in the Mistral family), appear to be more performant than larger proprietary ones. These models seem to reach an optimal compromise on number of parameters and overall performance where additional parameters could disrupt this equilibrium, leading to potentially over-fitting or plateauing performance [61].

Another interesting result of our analysis was that overall general-purpose models outperform both domain-specific models specialized in the biomedical domain and the one specialized for text summarization, regardless of model size. A possible explanation for this behavior might be that domain-specific models fine-tuned on biomedical text might be better for learning and understanding the complex biomedical terminology or lexical patterns but fail in summarization tasks. On the other hand, models specifically designed for text summarization might be good at summarizing in general but fail at capturing the complex biomedical meaning. That is why generalist models, leveraging their broad knowledge, seem to perform better [62]. Additionally, domain-specific models can “forget” the general knowledge that was acquired during the pre-training phase, experiencing a phenomenon called “catastrophic forgetting”, which represents an issue when the task requires both domain-specific knowledge, biomedical knowledge, and context understanding for text summarization [63].

Most reasoning-oriented models ranked in the middle, indicating moderate performance. This can be explained by the intrinsic multi-step logical reasoning nature of these models. While it can be advantageous for tasks that require breaking problems down into sequential steps like for mathematical operations or computer programming, it may not be ideal for text summarization which requires semantic compression and factual grounding instead [64].

Even though our results are based on a robust evaluation framework, there are several factors worth discussing. Model access methods varied across the evaluation due to differing Application Programming Interface (API) capabilities and requirements. HuggingFace models were accessed through their supported interfaces: the pipeline API (task="summarization") where available, or chat/completion formats for models that did not support the pipeline approach. Ollama models required use of the generate endpoint with merged prompts, while OpenAI, Anthropic, and Mistral models each mandated their respective provider-specific APIs (responses.create, messages.create, and chat.complete) with distinct message structures. We applied hyperparameter normalization where possible, though API-level constraints prevented full standardization. For example, GPT-5 does not support temperature control, instead offering only reasoning-specific parameters. Additionally, proprietary middleware layers may transform requests and responses in undocumented ways, potentially affecting outputs independently of the underlying model architectures. These necessary methodological variations warrant consideration when interpreting performance differences across models.

A limitation of this benchmarking study is that we focused on a single summarization task: generating concise summaries from biomedical abstracts. This setup provides a clear and well-defined evaluation framework, but the findings may not fully extend to other forms of scientific or biomedical summarization, including full-length articles, clinical trial data, or lay-oriented summaries. Given the rapid evolution of LLMs, these results just capture a specific snapshot in time and may change as newer architectures and models become available.

Another limitation lies in the exclusive reliance on au-

tomatic evaluation metrics. Although combining lexical-based, semantic-based, and factual consistency measures offers a broad view, human assessment could provide a more nuanced understanding of readability, coherence, and factual correctness. Future work could therefore extend this benchmark by integrating expert-based evaluations, exploring alternative summarization tasks, and including emerging model families as they are released.

The results of this benchmark provide useful guidance for selecting summarization models in biomedical and scientific settings. The strong performance of general-purpose LMs indicates that broad, diverse pretraining is often more advantageous than narrow domain adaptation when dealing with unseen scientific content.

Another key consideration is the trade-off between output quality and processing efficiency. While LLMs achieved the highest overall scores, SLMs deliver competitive results at substantially lower computational cost [65], which makes them especially attractive for large-scale or resource-constrained applications. Choosing between large and small models therefore depends not only on desired output quality but also on the intended scale of summarization.

Overall, the findings suggest that general-purpose LMs currently offer the most reliable and practical choice for biomedical text summarization. Their consistent performance across evaluation criteria demonstrates that broad generalization outweighs the marginal gains from more narrowly specialized or fine-tuned approaches, many of which are not primarily optimized for text summarization.

Acknowledgments

Enrico Bono is supported by a grant from the European Union's Horizon Europe Marie Skłodowska-Curie Actions Doctoral Networks program project PICKED (HORIZON–MSCA–2023–DN-01, grant number 101168626). Louiza Galou is supported by a grant from the European Union's Horizon Europe Marie Skłodowska-Curie Actions Doctoral Networks Industrial Doctorates program project PROMOTE (HORIZON–MSCA–2023–DN-01, grant number 101169245). Paul Perco and Matthias Ley are members of and would like to cordially thank the COST Action PerMediK, CA21165, supported by COST (European Cooperation in Science and Technology). We gratefully acknowledge Dorota Wojenska for her support in optimizing the overview workflow figure (Figure 1).

Author Contributions

Conceptualization: M.L., P.P., E.B. and F.B.; Methodology: M.L., P.P., E.B. and F.B.; Software: M.L.; Formal analysis: F.B., E.B., P.P. and M.L.; Validation: E.B., F.B., L.F., P.P., M.L., L.G., K.K.-I., S.W., P.A. and K.K.; Writing—original draft: F.B., E.B., M.L. and P.P.; Writing—review & editing: E.B., F.B., P.P., M.L., K.K., L.G., L.F., K.K.-I., P.A. and S.W.; Visualization: F.B. and E.B.; Supervision: M.L. and P.P.; All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

K.K. is co-founder and shareholder of Delta4 GmbH (Vienna, Austria). F.B., E.B., L.F., L.G., K.K.-I., S.W., P.A., P.P. and M.L. are employees of Delta4 GmbH (Vienna, Austria).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Anthropic (Claude 4.5 Opus) and OpenAI (ChatGPT-5) in order to enhance textual clarity and readability without introducing of new hypotheses or data. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Bibliography

1. Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, 2025. URL https://arxiv.org/abs/2403.02901.
2. Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. A systematic survey of text summarization: From statistical methods to large language models, 2024. URL https://arxiv.org/abs/2406.11289.
3. Mukesh Kumar Rohil and Varun Magotra. An exploratory study of automatic text summarization in biomedical and healthcare domain. *Healthcare Analytics*, 2:100058, 2022. ISSN 2772-4425. doi: https://doi.org/10.1016/j.health.2022.100058. URL https://www.sciencedirect.com/science/article/pii/S2772442522000223.
4. Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. A survey for biomedical text summarization: From pre-trained to large language models, 2023. URL https://arxiv.org/abs/2304.08763.
5. Hans Peter Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2: 159–165, 1958. URL https://api.semanticscholar.org/CorpusID:15475171.
6. H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, April 1969. ISSN 0004-5411. doi: 10.1145/321510.321519. URL https://doi.org/10.1145/321510.321519.
7. Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation - J DOC*, 60:503–520, 10 2004. doi: 10.1108/00220410410560582.
8. Lawrence H. Reeve, Hyoil Han, Saya V. Nagori, Jonathan C. Yang, Tamara A. Schwimmer, and Ari D. Brooks. Concept frequency distribution in biomedical text summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 604–611, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595934332. doi: 10.1145/1183614.1183701. URL https://doi.org/10.1145/1183614.1183701.
9. Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3252/.
10. Muhammad Afzal, Fakhare Alam, Khalid Mahmood Malik, and Ghaus M Malik. Clinical context–aware biomedical text summarization using deep neural network: Model development and validation. *J Med Internet Res*, 22(10):e19810, Oct 2020. ISSN 1438-8871. doi: 10.2196/19810. URL http://www.jmir.org/2020/10/e19810/.
11. Ahmed Almasoud, Siwar Hassine, Fahd Al-Wesabi, Mohamed Nour, Anwer Hilal, Mesfer Al Duhayyim, Ahmed Hamza, and Abdelwahed Motwakel. Automated multi-document biomedical text summarization using deep learning model. *Computers, Materials & Continua*, 71:5800, 01 2022. doi: 10.32604/cmc.2022.024556.
12. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.
13. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

14. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019. URL http://arxiv.org/abs/1910.13461.
15. Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. BioBART: Pretraining and evaluation of a biomedical generative language model. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.9. URL https://aclanthology.org/2022.bionlp-1.9/.
16. S Abinaya, M.s.Antony Vigil, K Keerthika, and R Varshasri. Medical text summarization using bart with lora-based parameter efficient fine tuning. 01 2024.
17. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.
18. Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020. URL https://arxiv.org/abs/1912.08777.
19. Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.
20. Oleksandr Steblianko, Volodymyr Shymkovych, Peter Kravets, Anatolii Novatskyi, and Lyubov Shymkovych. Scientific article summarization model with unbounded input length. *Information, Computing and Intelligent systems*, pages 150–158, 12 2024. doi: 10.20535/2786-8729.5.2024.314724.
21. Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Multi-step reasoning with large language models, a survey, 2025. URL https://arxiv.org/abs/2407.11511.
22. Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.
23. Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.
24. Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.
25. Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.
26. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.
27. Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.
28. Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.
29. Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.
30. Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 09 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409. bbac409.
31. Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, et al. Granite code models: A family of open foundation models for code intelligence, 2024. URL https://arxiv.org/abs/2405.04324.
32. Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.
33. Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, et al. Magistral, 2025. URL https://arxiv.org/abs/2506.10910.
34. Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.
35. Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.
36. Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. Scilitllm: How to adapt llms for scientific literature understanding, 2025. URL https://arxiv.org/abs/2408.15545.
37. Chengen Wang and Murat Kantarcioglu. A review of deepseek models' key innovative techniques, 2025. URL https://arxiv.org/abs/2503.11486.
38. Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, et al. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. https://arxiv.org/abs/2509.14233, 2025.
39. Elsevier. Highlights, 2024. https://www.elsevier.com/researcher/author/tools-and-resources/highlights (accessed: 2025-08-07).
40. Cell Press. Final submission: Other components: Highlights, 2024. https://www.cell.com/cell/information-for-authors/final-submission (accessed: 2025-08-07).
41. Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai, 2025. URL https://arxiv.org/abs/2506.02153.
42. Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.
43. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.
44. Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.
45. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.
46. Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.
47. Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. URL https://arxiv.org/abs/2004.09297.
48. Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function, 2023. URL https://arxiv.org/abs/2305.16739.
49. Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021. URL https://arxiv.org/abs/2007.12626.
50. Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.454. URL https://aclanthology.org/2020.acl-main.454/.
51. Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.
52. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
53. Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
54. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. ISSN 0169-7552. doi: https://doi.org/10.1016/S0169-7552(98)00110-X. URL https://www.sciencedirect.com/science/article/pii/S016975529800110X. Proceedings of the Seventh International World Wide Web Conference.
55. Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.
56. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.
57. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.
58. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.
59. Zhicheng Wang, Sahil Kulkarni, Jia Nong, Marco Zamora, Alireza Ebrahimimojarad, Elizabeth Hood, Tea Shuvaeva, Michael Zaleski, Damodar Gullipalli, Emily Wolfe, et al. A percolation phase transition controls complement protein coating of surfaces. *Cell*, 188(15):4058–4073.e25, 2025. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2025.05.026. URL https://www.sciencedirect.com/science/article/pii/S0092867425005768.
60. Borui Xu, Yao Chen, Zeyi Wen, Weiguo Liu, and Bingsheng He. Evaluating small language models for news summarization: Implications and factors influencing performance, 2025. URL https://arxiv.org/abs/2502.00641.
61. Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2025. URL https://arxiv.org/abs/2305.16264.
62. Felix J. Dorfner, Amin Dada, Felix Busch, Marcus R. Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C. Adams, and Keno K. Bressem. Biomedical large languages models seem not to be superior to generalist models on unseen medical data, 2024. URL https://arxiv.org/abs/2408.13833.

63. Yuxuan Zhou, Xien Liu, Xiao Zhang, Chen Ning, Shijin Wang, Guoping Hu, and Ji Wu. Investigating and mitigating catastrophic forgetting in medical knowledge injection through internal knowledge augmentation learning. *OpenReview*, 2025. Preprint. Available online: https://openreview.net/forum?id=i9RDDi2SZC.
64. Keyan Jin, Yapeng Wang, Leonel Santos, Tao Fang, Xu Yang, Sio Kei Im, and Hugo Gonçalo Oliveira. Reasoning or not? a comprehensive evaluation of reasoning llms for dialogue summarization, 2025. URL https://arxiv.org/abs/2507.02145.
65. Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Jayanaka Dantanarayana, Krisztian Flautner, Lingjia Tang, Yiping Kang, and Jason Mars. Scaling down to scale up: A cost-benefit analysis of replacing openai's llm with open source slms in production, 2024. URL https://arxiv.org/abs/2312.14972.

Supplementary Information

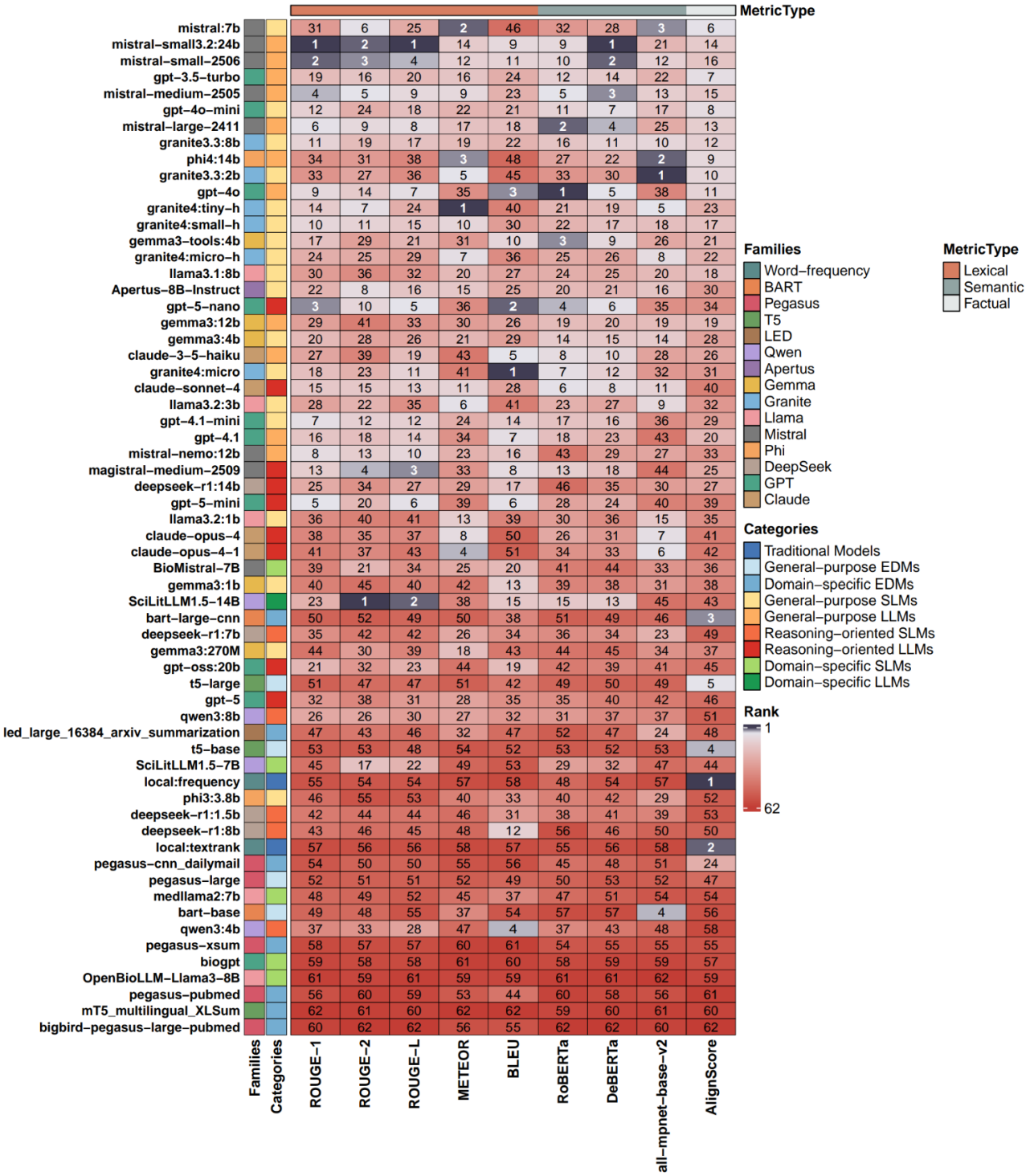

Figure S1. Model performance across metrics: Overview of the performance of all evaluated models across all metrics.